Photorealistic Video Generation with Diffusion Models

Year: 2023 | Citations: 256 | Authors: Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn

---

Abstract

We present W.A.L.T, a transformer-based approach for photorealistic video generation via diffusion modeling. Our approach has two key design decisions. First, we use a causal encoder to jointly compress images and videos within a unified latent space, enabling training and generation across modalities. Second, for memory and training efficiency, we use a window attention architecture tailored for joint spatial and spatiotemporal generative modeling. Taken together these design decisions enable us to achieve state-of-the-art performance on established video (UCF-101 and Kinetics-600) and image (ImageNet) generation benchmarks without using classifier free guidance. Finally, we also train a cascade of three models for the task of text-to-video generation consisting of a base latent video diffusion model, and two video super-resolution diffusion models to generate videos of 512×896 resolution at 8 frames per second.